b-jet Identification in the D0 Experiment

Sébastien Greder

Institut Pluridisciplinaire Hubert Curien, Strasbourg

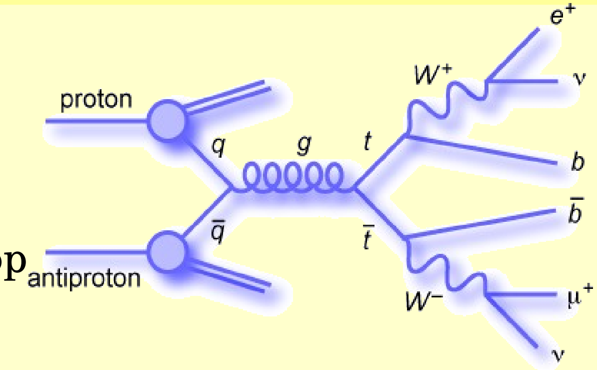
Outline

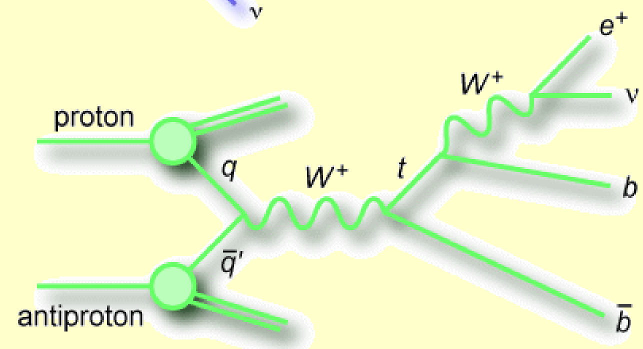
- Introduction
- The DØ detector
- Algorithms
- Performance measurement
  - B-jet efficiency
  - Fake tag rate
- Conclusion

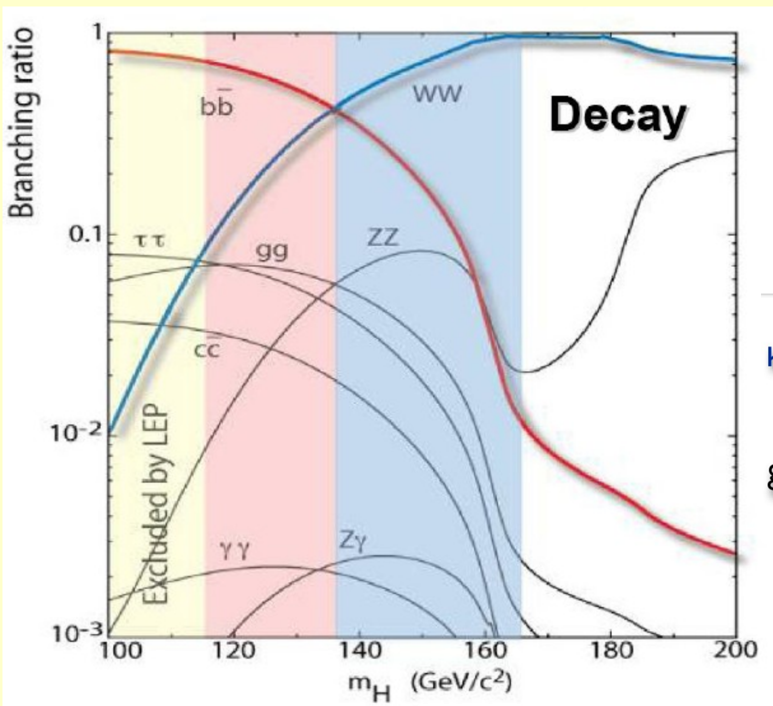
Introduction

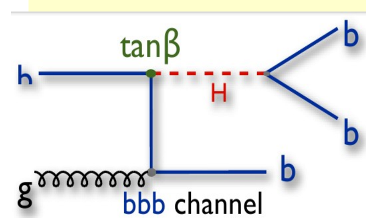
Physics

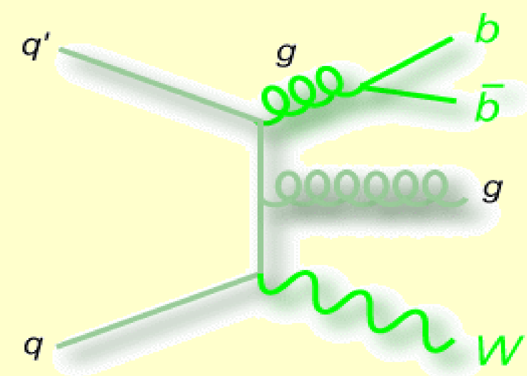
- Top physics: x-section, mass, single-top
- "Backgrounds": W/Z+heavy flavour
- Higgs searches: Low-mass, SUSY





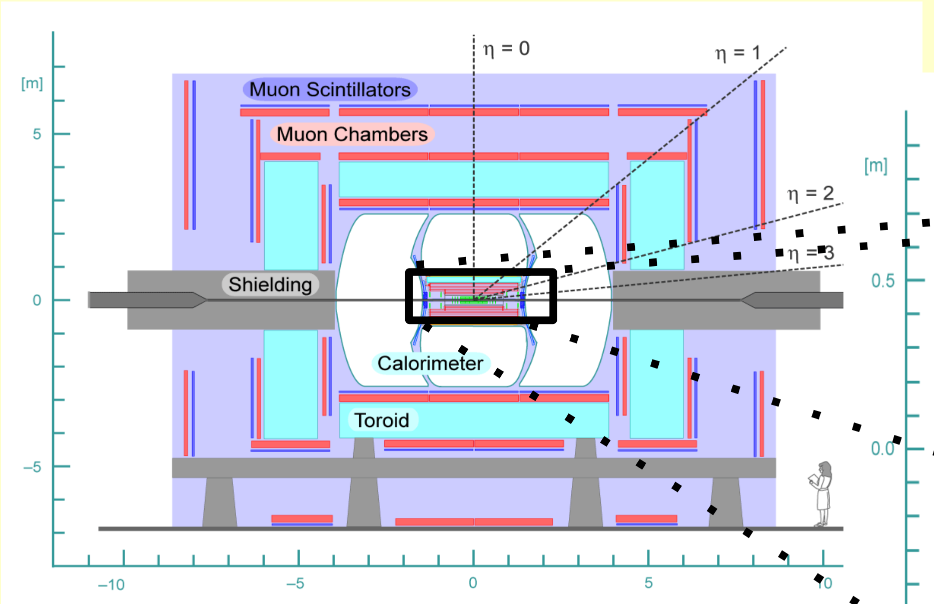


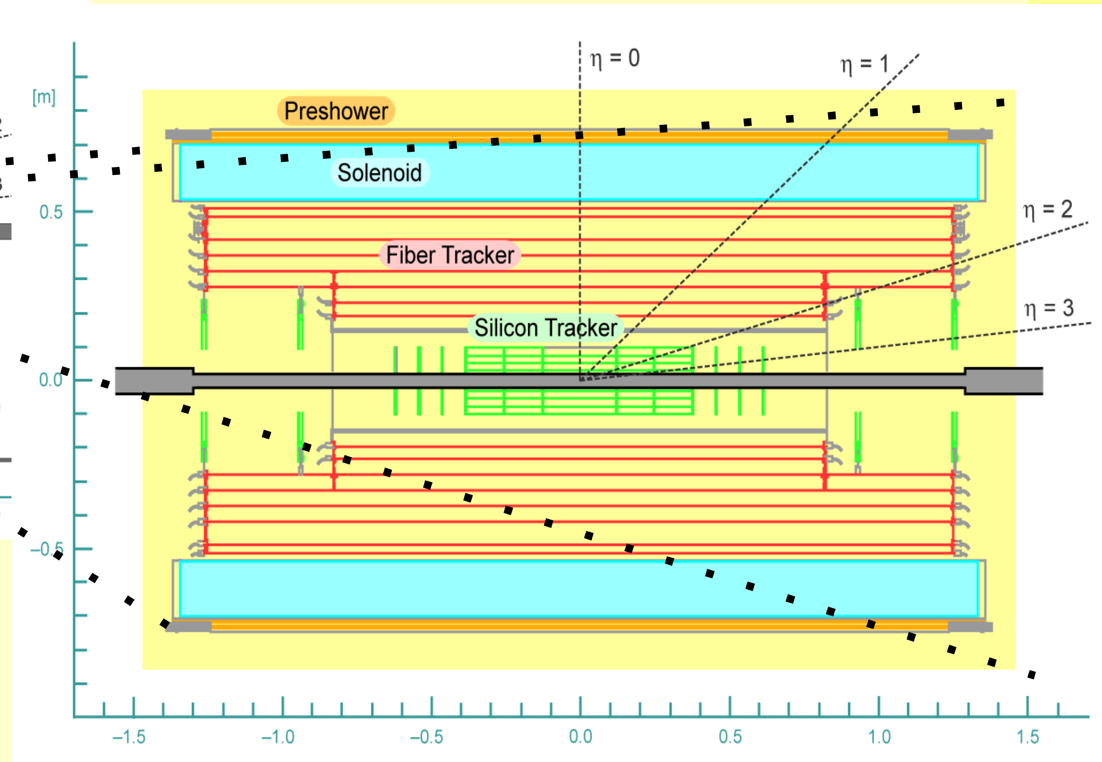




Sébastien Greder

The DØ detector





Silicon tracker (SMT)

- http://arxiv.org/abs/1005.0801
- 6 barrels, 4 layers each, z ~1 m
  + new Layer 0 @ r =1.6 cm (RunIIb, see: http://arxiv.org/abs/0911.2522)
- Coverage $|\eta| < 2.5$

Central Fiber Tracker (CFT)

- 8 layers of scintillating fiber (axial and stereo)
- 20 < r < 51 cm in **2T magnetic field**

Muon system covers $|\eta|<2$

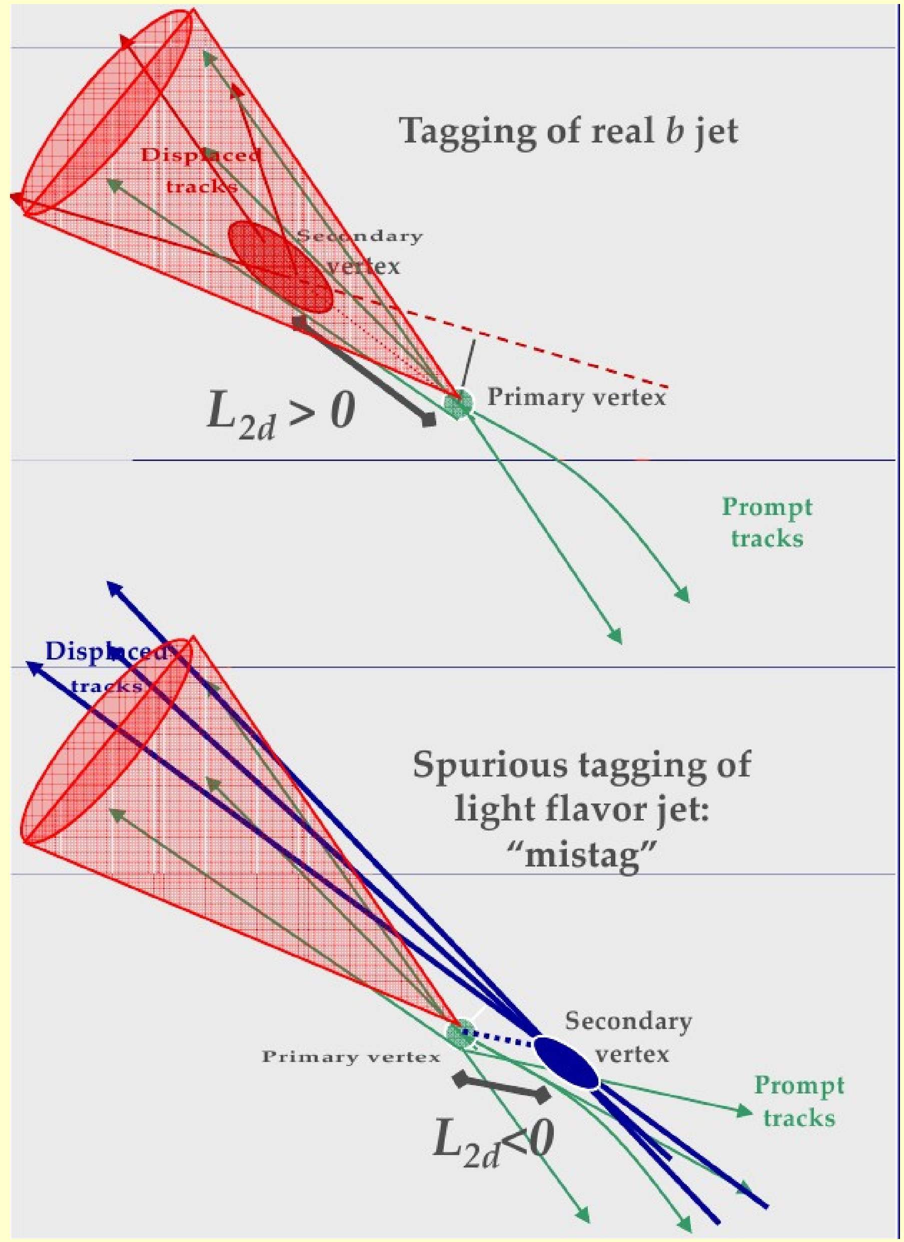
Sébastien Greder

Institut Pluridisciplinaire Hubert Curien STRASBOURG

Tag vs. mis-tags

B hadrons properties

- Mass: ~5 GeV/c^2
- Decay length: ~3mm
- **Hard** fragmentation
- Semi-leptonic decays



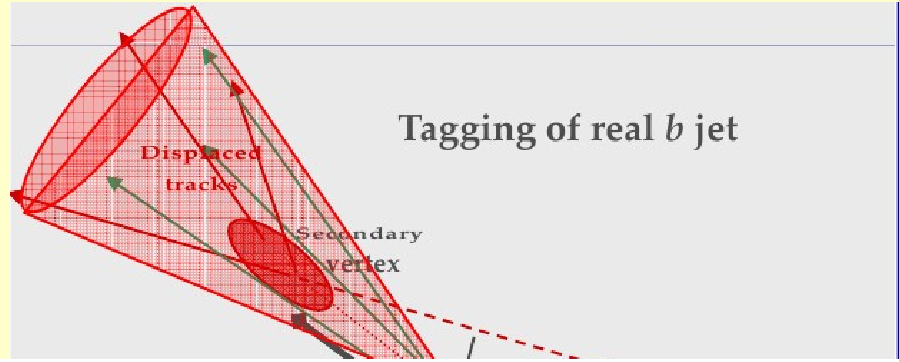
Fake / Mis -tags

- Primary vertex resolution
- Track parameters resolutions
- Long lived particles
- Secondary interactions

Tag vs. mis-tags

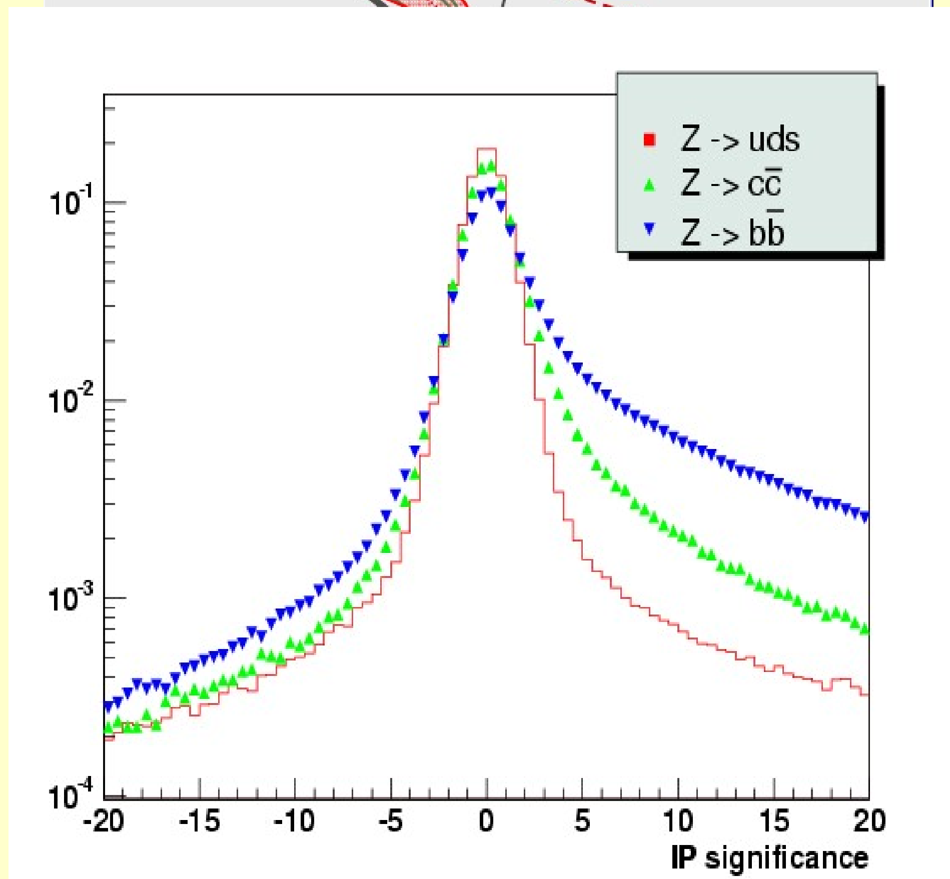
B hadrons properties

- Mass: ~5 GeV/c^2
- Decay length: ~3mm
- **Hard** fragmentation
- Semi-leptonic decays



Fake / Mis -tags

- Primary vertex resolution
- Track parameters' resolution
- Long lived particles
- Secondary interactions



Sébastien Greder

IPHC Institut Pluridisciplinaire Hubert Curien STRASBOURG

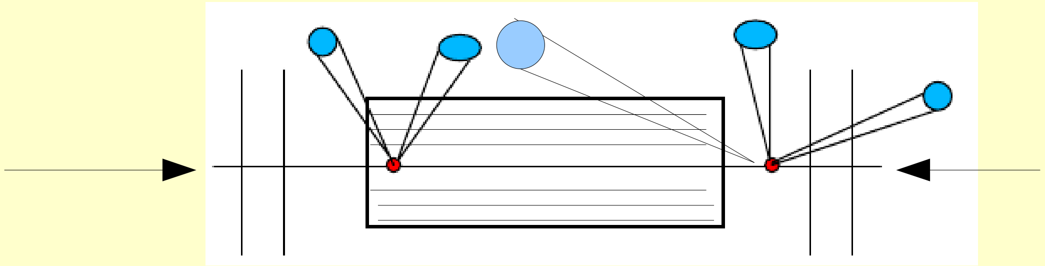
Tagging prerequisites

Taggability

In the following tagging algorithms are only based on **tracking and vertexing** of charged particles

↳ only (calorimeter) jets with minimum tracking information are considered

- Interaction region, $\sigma_z \approx 25$cm, + detector acceptance affect track reconstruction efficiencies
  ↳ performance dependence on η and interaction point's Z coordinate.



- Fraction of fake jets is ~small, but depends on **final** state.
  ↳ Decoupling this effect from the tagging algorithms properly allows the extraction of a tagging performance which can be assumed to be **universal**, i.e., applicable to *any general final states*

Tagging prerequisites (II)

Taggable jets are thus defined as follow:

- **2-step clustering:**

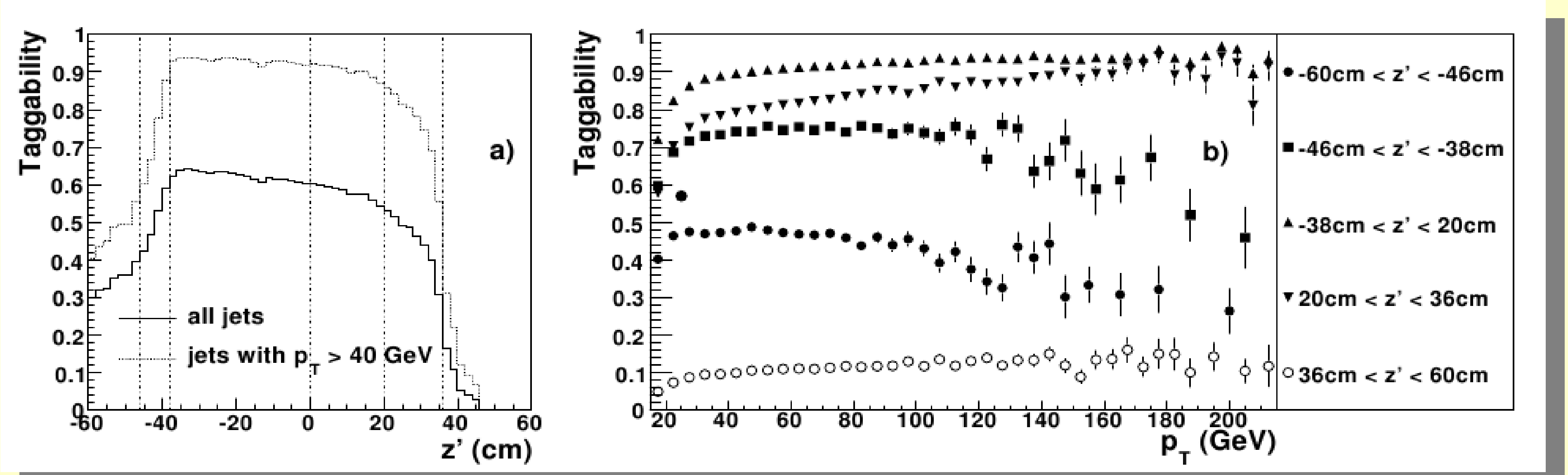
  i. along beam axis (dca_Z < 4 mm)

  ii. 0.5 cone (snow mass) jets *(within each z-cluster)*

  ↳ finally require: ΔR(*calo-jet,track-jet*) < 0.5

- **Track-jets:** 1 SMT hit tracks, seed track pT > 1GeV/c, pT > 0.5 GeV/c for other tracks

- **Parametrized as:** $F(p_T, \eta, z')$, with $z' \equiv |z| \cdot \mathrm{sign}(\eta \cdot z)$



Sébastien Greder

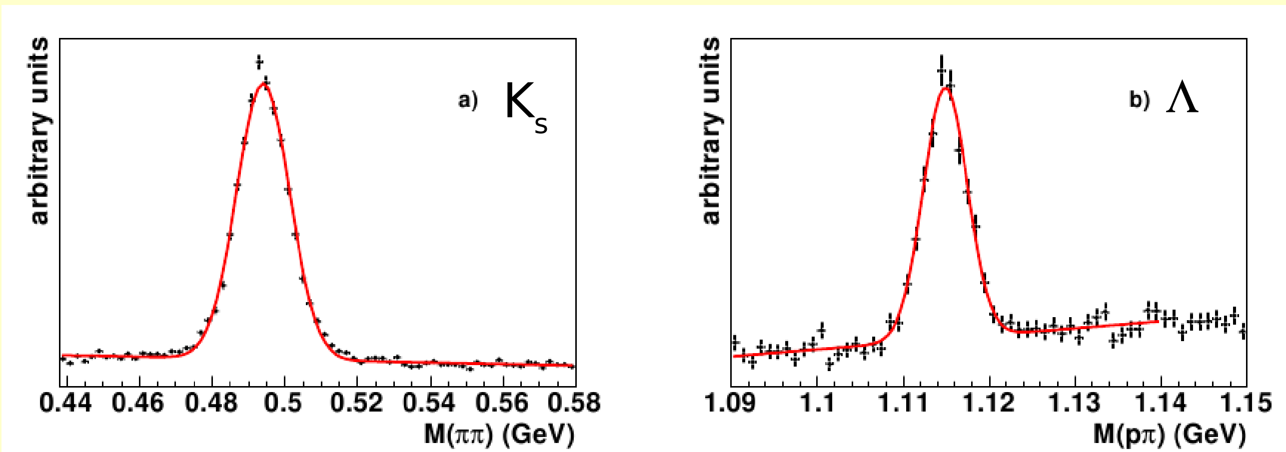
Tagging prerequisites (III)

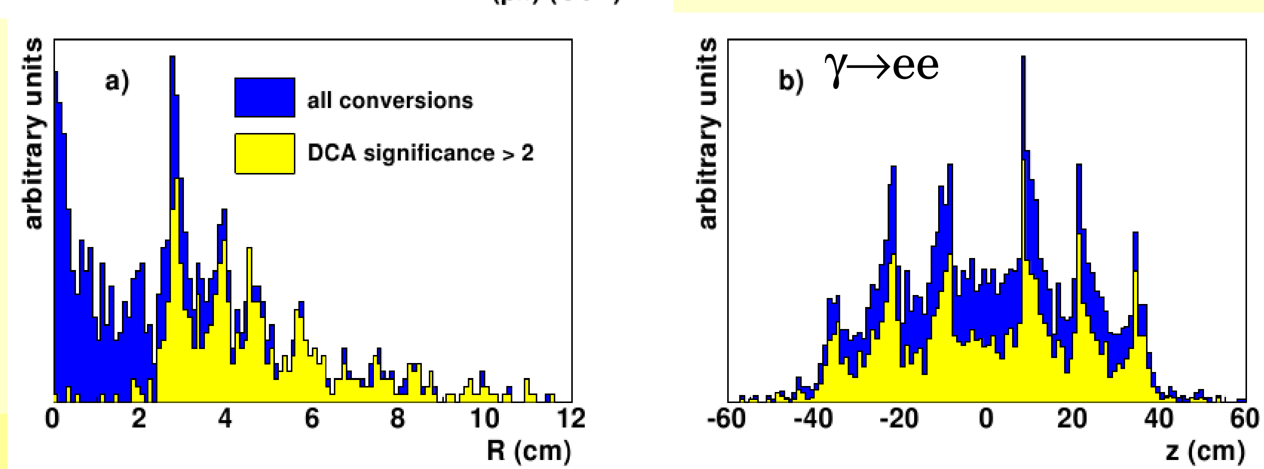
Track preselection

- Each b-id. algorithm uses its own track reconstruction quality criteria

V^0 removal

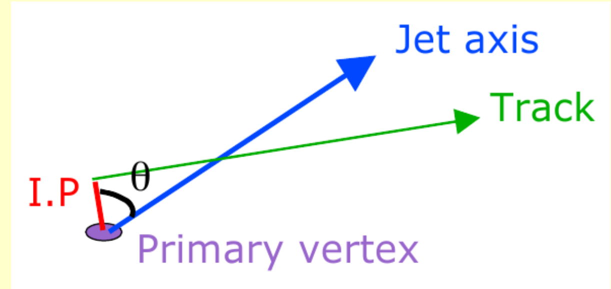
- Light strange hadrons have long lifetimes
- Photon conversions can occur at large distances in the tracker material



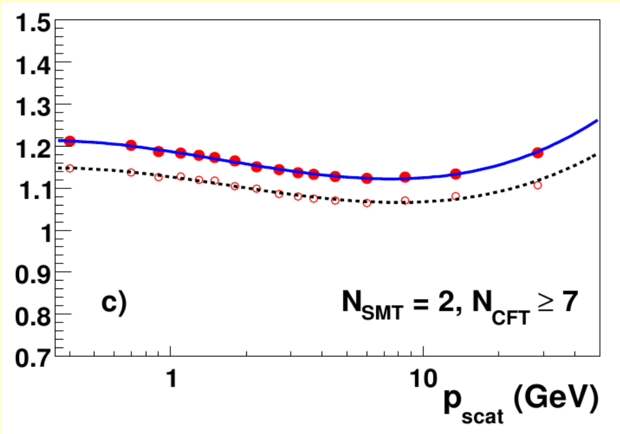


Algorithms (I)

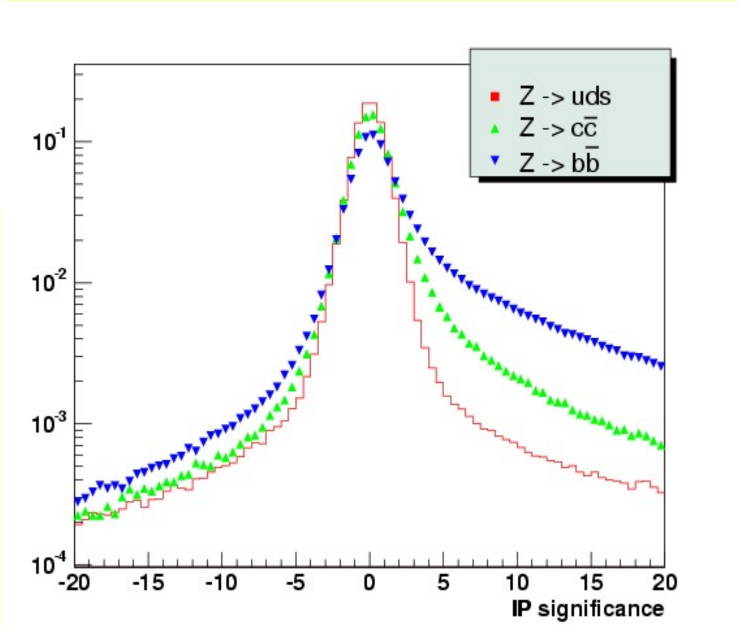
Impact Parameter (IP) based tagger



- IP and its significance S_{IP} are **signed** w.r.t jet direction
- IP error calibrated in data and simulation for ***multiple-scattering effects*** and ***PV resolution dependence***



$p_{scat} = p \sin\theta^{3/2}$



Discrete (CSIP)

- counts tracks with: $|S_{IP}|$ > cut *(2 >3 | | 3 > 2)*

Continuous (JLIP)

- p.d.f from negative IP resolution function, ***R(s)***

Algorithms (II)

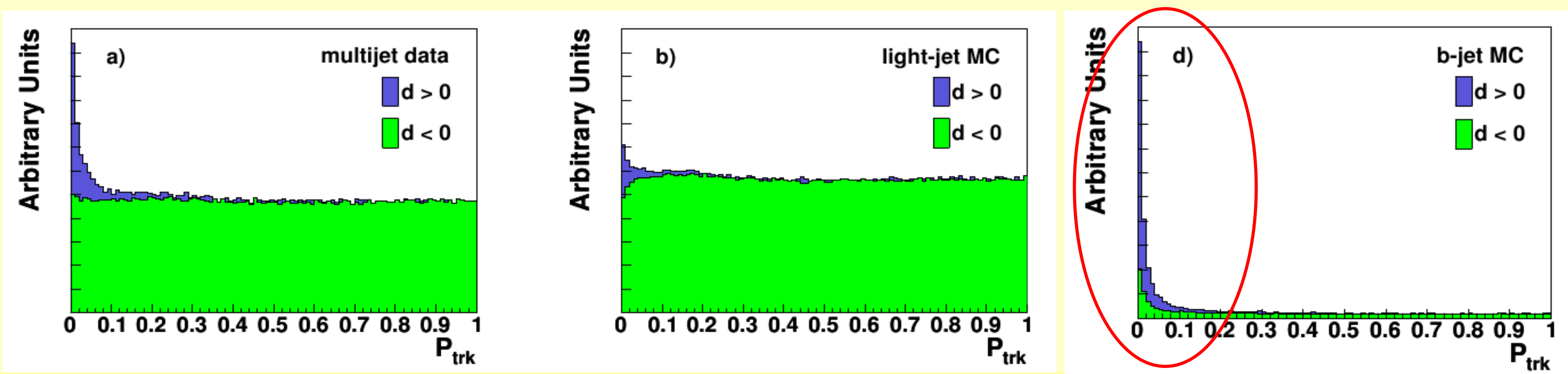
Impact Parameter (IP) based tagger

- **Continuous (JLIP)**
  - p.d.f from negative IP resolution function, ***R(s)***

$$\mathcal{P}_{\rm trk}(\mathcal{S}_d^{\rm corr}) = \frac{\int_{-\infty}^{-|\mathcal{S}_d^{\rm corr}|} \mathcal{R}(s)ds}{\int_{-\infty}^{0} \mathcal{R}(s)ds}$$

$$\mathcal{P}_{jet}^{\pm} = \Pi^{\pm} \times \sum_{j=0}^{N_{trk}^{\pm}-1} \frac{(-\log \Pi^{\pm})^j}{j!} \quad \text{with} \quad \Pi^{\pm} = \prod_{i=1}^{N_{trk}^{\pm}} \mathcal{P}_{trk}(\mathcal{S}_{IP<0}^{IP>0})$$

***Note:** one can consider any set of tracks and e.g build an "hemisphere-probability" (Z-->bb, LEP)*



Sébastien Greder

IPHC Institut Pluridisciplinaire Hubert Curien Strasbourg

Algorithms (II)

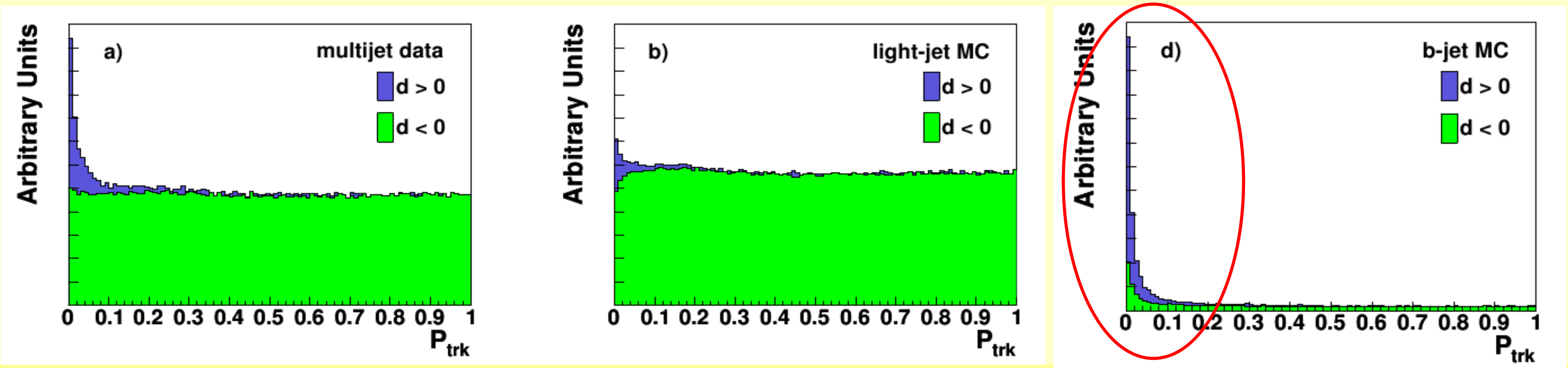
Impact Parameter (IP) based tagger

- **Continuous (JLIP)**
  - p.d.f from negative IP resolution function, ***R(s)***

$$\mathcal{P}_{\mathrm{trk}}(\mathcal{S}_d^{\mathrm{corr}}) = \frac{\int_{-\infty}^{-|\mathcal{S}_d^{\mathrm{corr}}|} \mathcal{R}(s)ds}{\int_{-\infty}^{0} \mathcal{R}(s)ds}$$

For TMVA aficionados, this is what is called "Rarity" in the TMVAGui.C

Note: *one can consider any set of tracks and e.g build an "hemisphere-probability" (Z-->bb, LEP)*



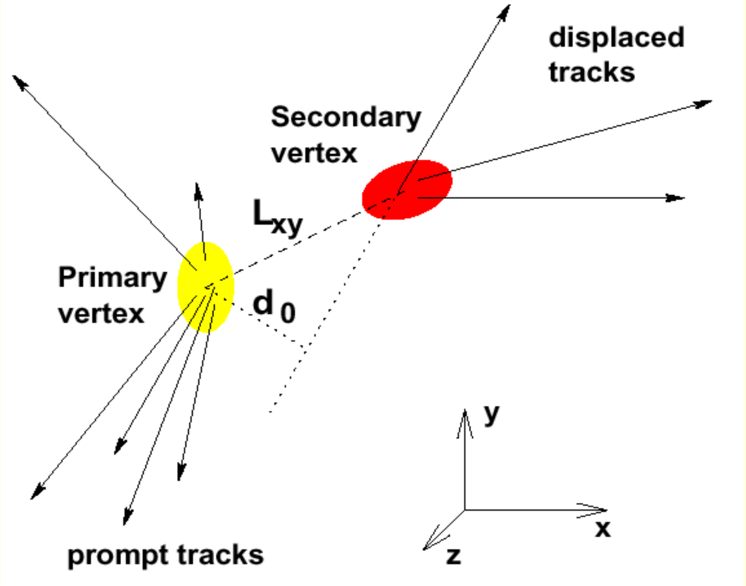
Sébastien Greder

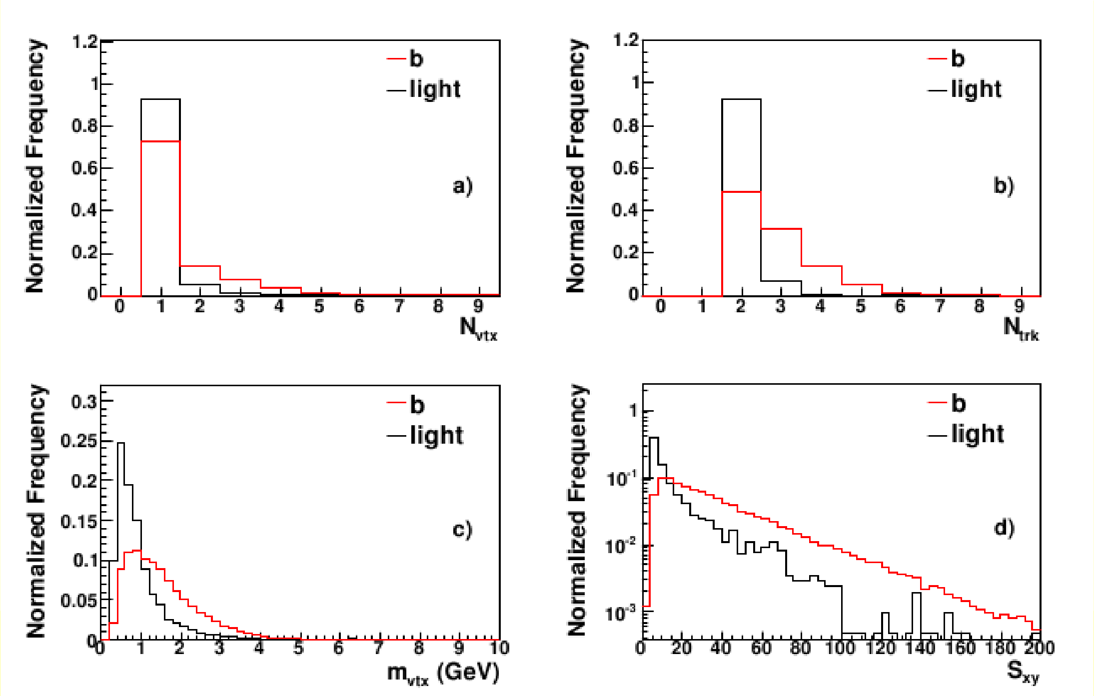
IPHC Institut Pluridisciplinaire Hubert Curien Strasbourg

Algorithms (III)

Secondary vertex, SVT

- Starts from track-based jets *(simple cone algo.)*
- Kalman-filter based vertex finder
- Track pruning w.r.t χ^2 contribution to vertex
- **Tag is defined if:**
  ΔR(vertex,jet) < 0.5 and if
  decay length significance, S_{Lxy} > cut



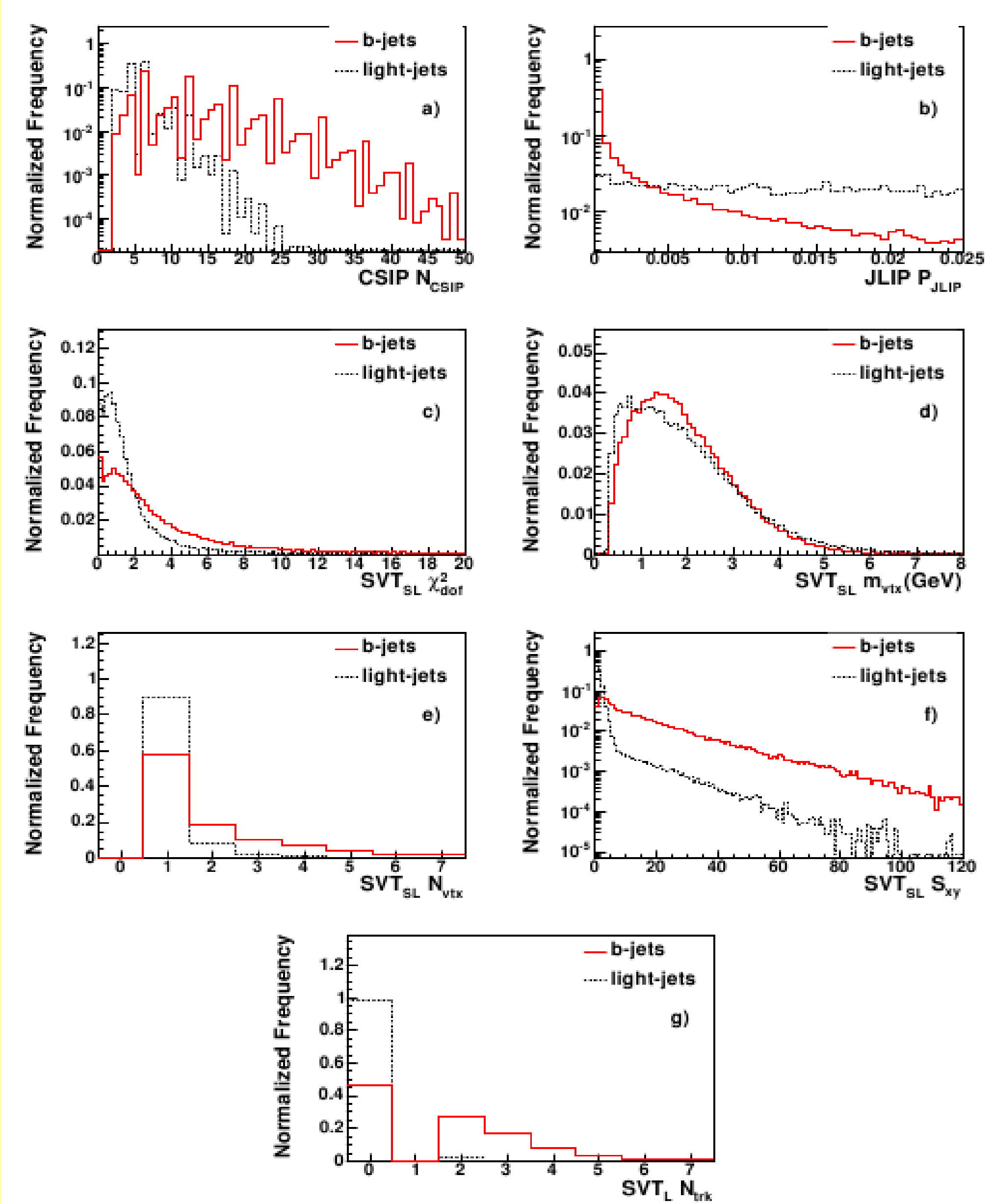


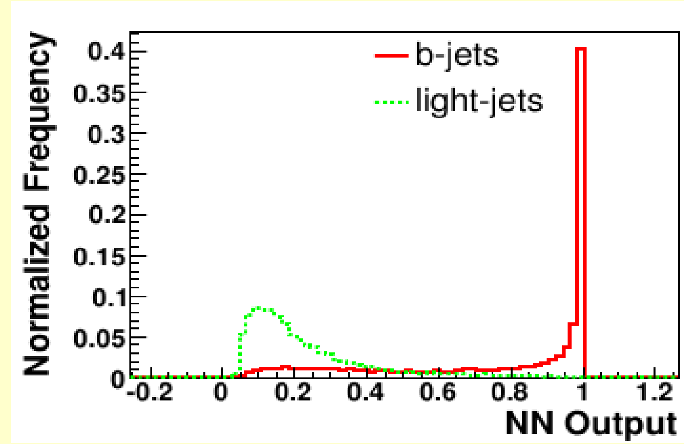
Sébastien Greder

Institut Pluridisciplinaire Hubert Curien Strasbourg

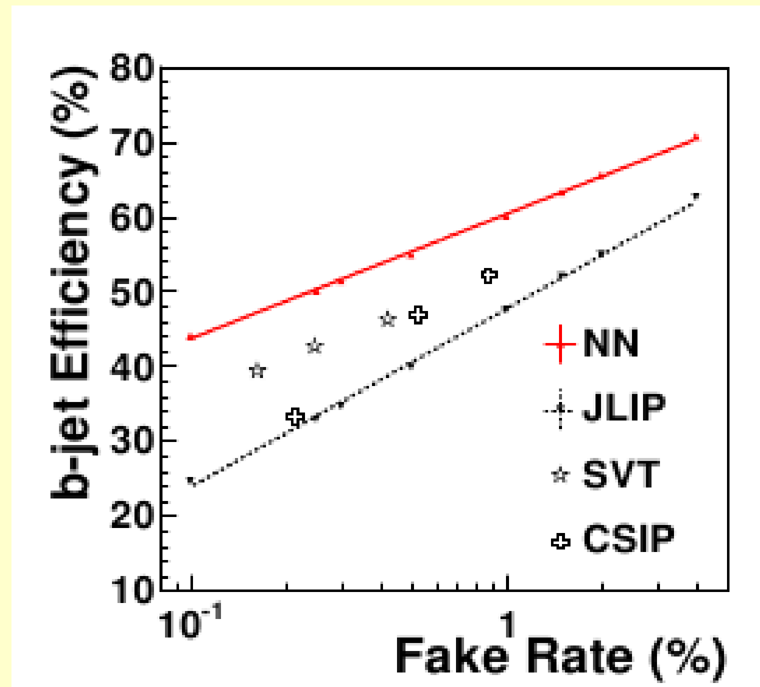
All in one: Neural Network tagger

Optimized selection of inputs: CSIP, JLIP & 5 SVT properties





... can lead to significant improvement:



Sébastien Greder

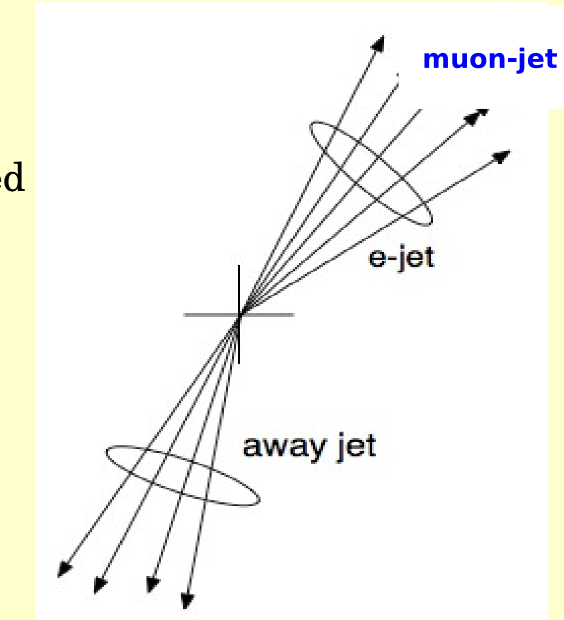
Institut Pluridisciplinaire Hubert CURIEN STRASBOURG

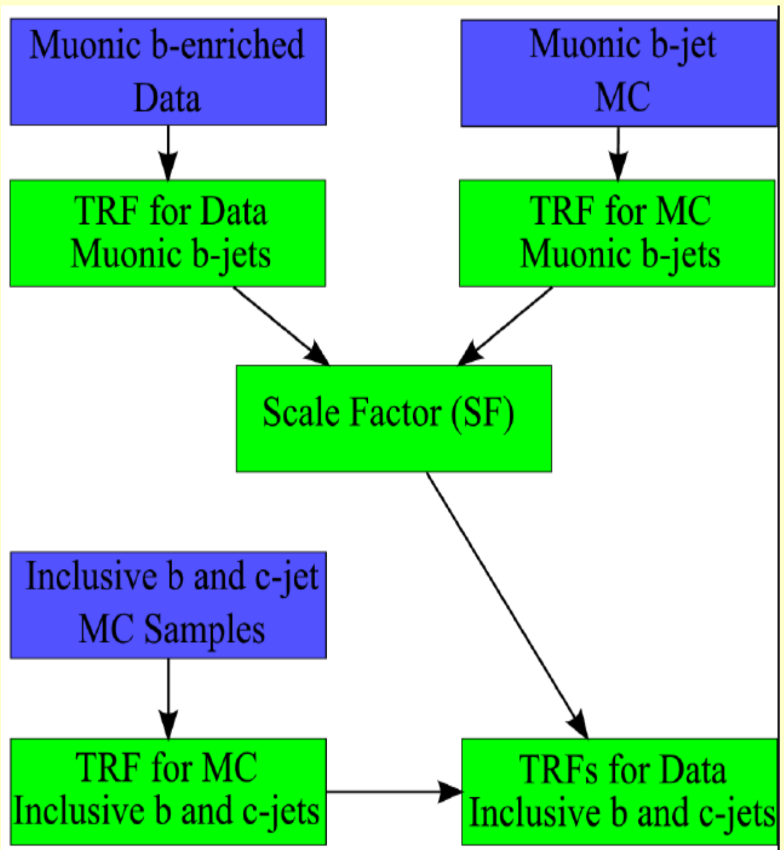
Performance measurements

B-identification efficiency (I)

Measured in data

- Using b-enriched data samples:
  Di-jet back-to-back sample & require $\Delta R(<0.5)$ matched soft (> 4GeV/c) muon in jet
- Efficiency extracted using **SystemD** method





- Muonic data/MC b-Scale Factor:
  $\mathbf{SF}_{b\rightarrow\mu}(\mathbf{p}_T,\eta)$
- Apply SF to inclusive b & c Tag Rate Function (TRF):

$$\varepsilon_b^{\mathrm{data}} = \frac{\varepsilon_{b\to\mu X}^{\mathrm{data}}\cdot\varepsilon_b^{\mathrm{MC}}}{\varepsilon_{b\to\mu X}^{\mathrm{MC}}} = \mathrm{SF}_b\cdot\varepsilon_b^{\mathrm{MC}}$$

B-identification efficiency (II)

The SystemD method

- Historically developed to measure efficiency *solely in data*
- Simulation only used for corrections factors (MC/MC ratios)
- **Main idea:** use *uncorrelated* selection criteria (i.e taggers) applied to various data samples and build a system of (non-linear) equations

General case:

- Consider **s** data samples composed of 1 signal and **f** backgrounds. Each sample **j** can gives ***1+f unknowns*:** the signal and backgrounds fractions :

$$n_{i=0\dots f}^{j=1\dots s} \quad \textit{constrained by:} \quad \sum_{i=0}^{f} n_i^j = 1$$

- Each tagger **k** gives also 1+f unknowns, the efficiencies : $\varepsilon_{i=0\dots f}^{k=1\dots t}$
- When applying the tagger k on sample j, only a fraction q_j^k of the total number of events survives:

$$q_j^k = \sum_{i=0}^{f} \varepsilon_i^k n_i^j$$

- When applying e.g 2 uncorrelated criteria: $$q_j^{k_1,k_2} = \sum_{i=0}^{f} \varepsilon_i^{k_1} \varepsilon_i^{k_2} n_i^j$$

Sébastien Greder

B-identification efficiency (III)

- Combining **t** taggers and **s** samples ⇒ **2^t.s** equations
- To solve the system, one needs at least as many equations as unknowns:

$$2^t \cdot s \geq (1+f)(s+t)$$

- The first combinations are:

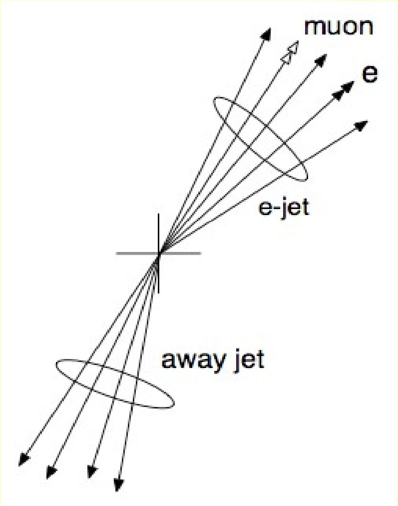
  - $s=2,\ t=2,\ f=1:$ 8 equations and 8 unknowns;
  - $s=1,\ t=3,\ f=1:$ 8 equations and 8 unknowns;
  - $s=2,\ t=3,\ f=2:$ 16 equations and 15 unknowns;
  - $s=6,\ t=2,\ f=2:$ 24 equations and 24 unknowns.

In practice finding many samples and (uncorrelated) taggers is <u>difficult</u>

<u>Note:</u> *t = 1, s = 2, f = 1 is known as the Matrix Method :-)*

B-identification efficiency (IV)

SystemD and b-tagging



s = 2 and only the first combination is considered:

- **2** *(uncorrelated)* **taggers:**

  NN-tagger & soft lepton(muon) tagger w/ a p_T^{rel} cut

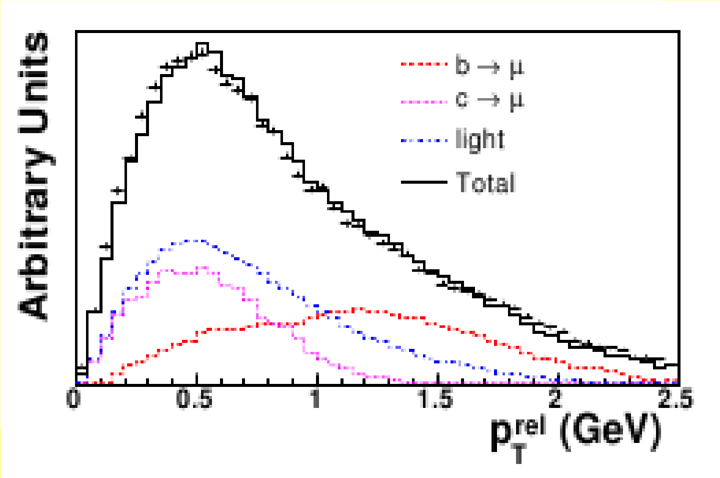
- **2 data samples** w/ different flavour content:

  muon-jet & muon-jet + away tag

- Apply 2 taggers **separately / simultaneously** on 2 samples and solve *(analytically or numerically)* the 8 equations / 8 unknowns among which:

  ε_b**(NN)**

- *c and light* jets are considered as a single background *(i.e f = 1)*



Sébastien Greder

IPHC Institut Pluridisciplinaire Hubert Curien Strasbourg

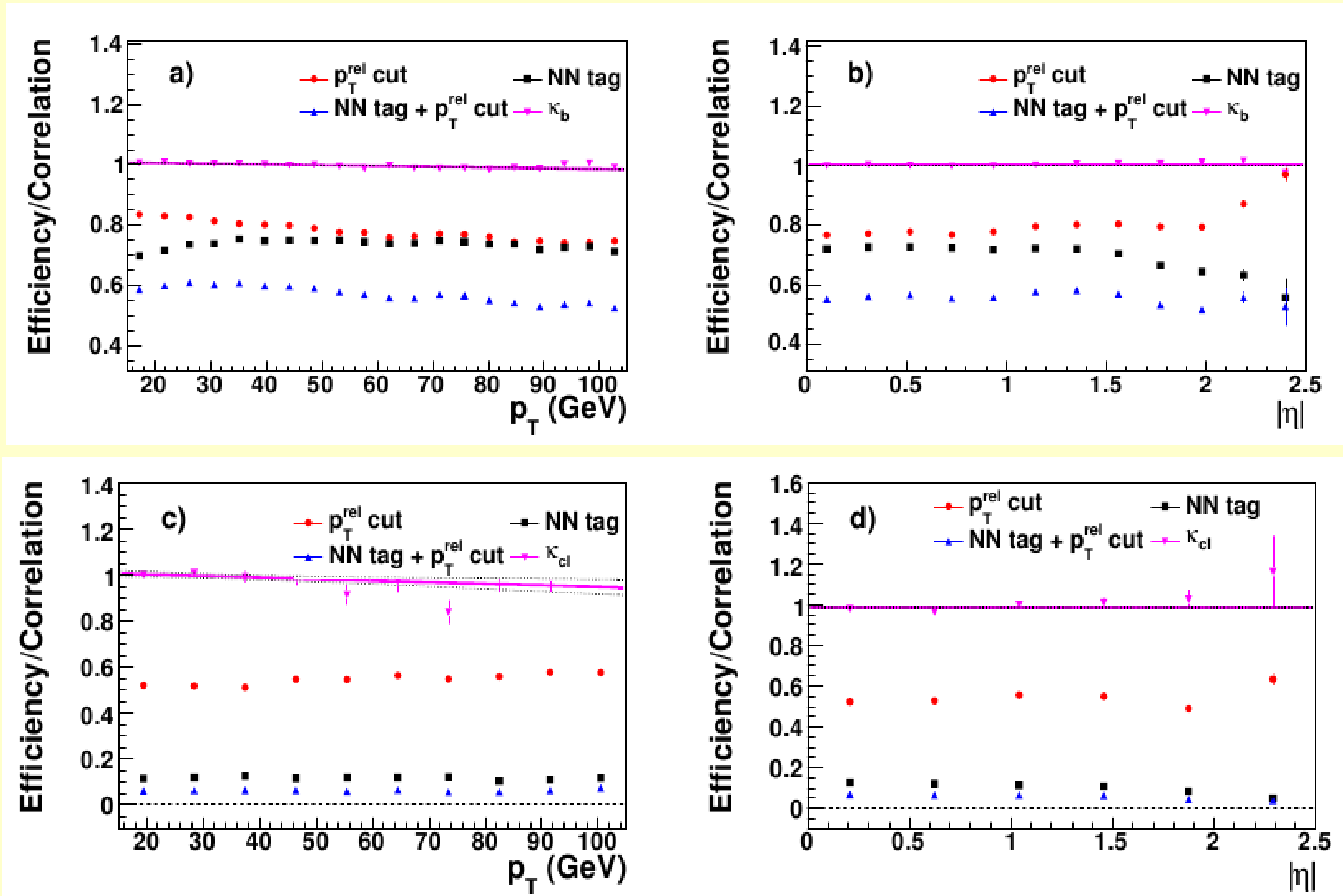
B-identification efficiency (V)

Corrections factors

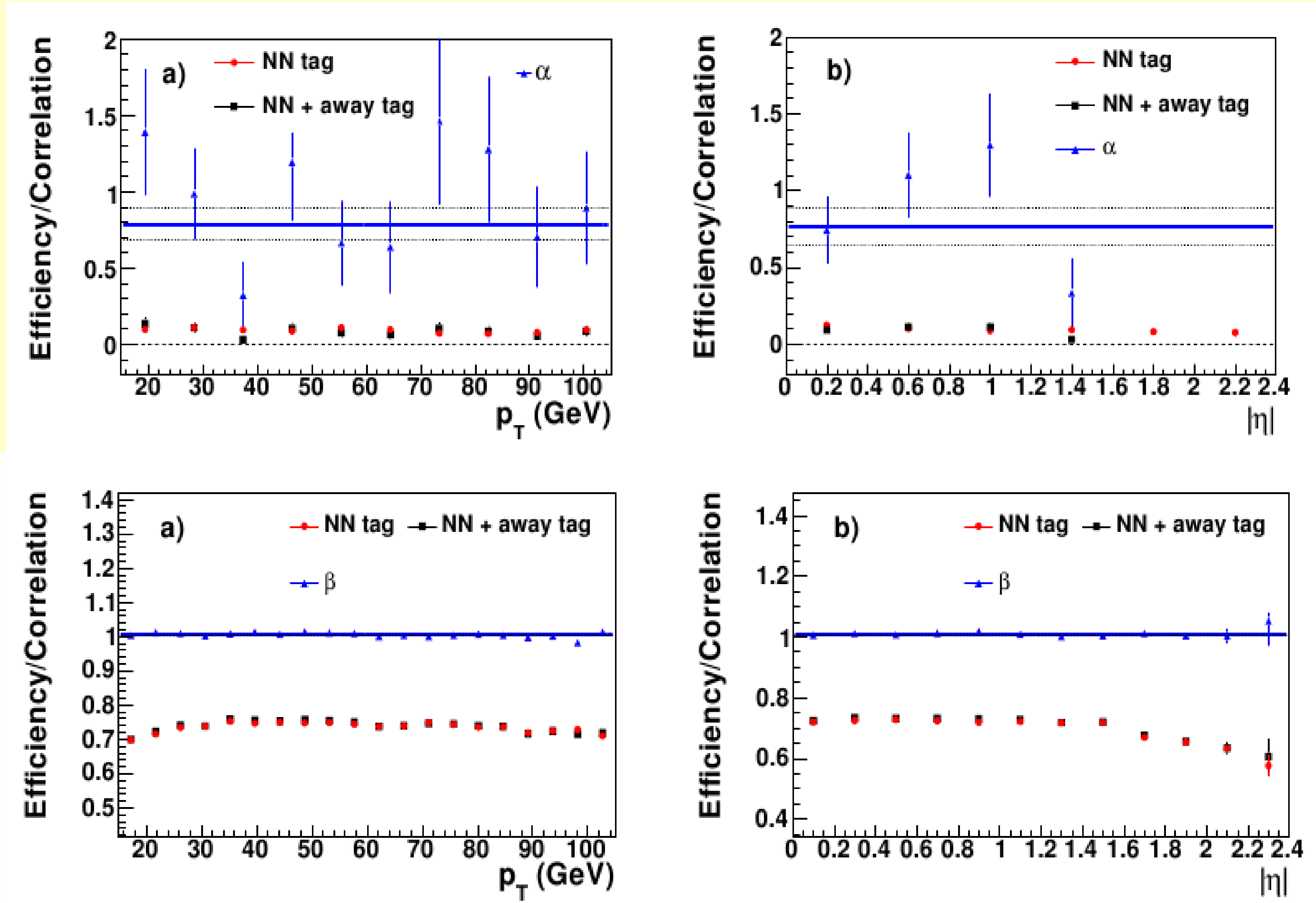
- The SLT and NN are assumed to be uncorrelated *(mass vs. lifetime)*
- The away-tag and the SLT are uncorrelated *(but same PV!)*
  - Introduce corrections factors for signal and backgrounds to quantify these correlations
  - Parameterized as a function of jet p_T/eta

$$
\begin{array}{lcccl}
f_b & + & f_{cl} & = & 1 \\
f_b \varepsilon_b^l & + & f_{cl} \varepsilon_{cl}^l & = & Q^l \\
f_b \varepsilon_b^m & + & f_{cl} \varepsilon_{cl}^m & = & Q^m \\
f_b \varepsilon_b^b & + & f_{cl} \varepsilon_{cl}^b & = & Q^b \\
f_b \kappa_b \varepsilon_b^l \varepsilon_b^m & + & f_{cl} \kappa_{cl} \varepsilon_{cl}^l \varepsilon_{cl}^m & = & Q^{l,m} \\
f_b \varepsilon_b^m \varepsilon_b^b & + & f_{cl} \varepsilon_{cl}^m \varepsilon_{cl}^b & = & Q^{m,b} \\
f_b \beta \varepsilon_b^b \varepsilon_b^l & + & f_{cl} \alpha \varepsilon_{cl}^b \varepsilon_{cl}^l & = & Q^{b,l} \\
f_b \kappa_b \beta \varepsilon_b^l \varepsilon_b^m \varepsilon_b^b & + & f_{cl} \kappa_{cl} \alpha \varepsilon_{cl}^l \varepsilon_{cl}^m \varepsilon_{cl}^b & = & Q^{l,m,b}
\end{array}
$$

B-identification efficiency (VI)



B-identification efficiency (VII)



Sébastien Greder

IPHC Institut Pluridisciplinaire Hubert Curien STRASBOURG

B-identification efficiency (VIII)

Systematic uncertainties

- Corrections factors measured with finite stat. MC
  ↳ vary within errors 1 correction *(fix the others)*
  and re-run SystemD
- p_T^{rel} cut varied from 0.3 to 0.8 GeV/c
- Add all errors quadratically
- Apply in *each jet p_T and eta* bins for *each operating point (OP)*

SystemD syst. errors:

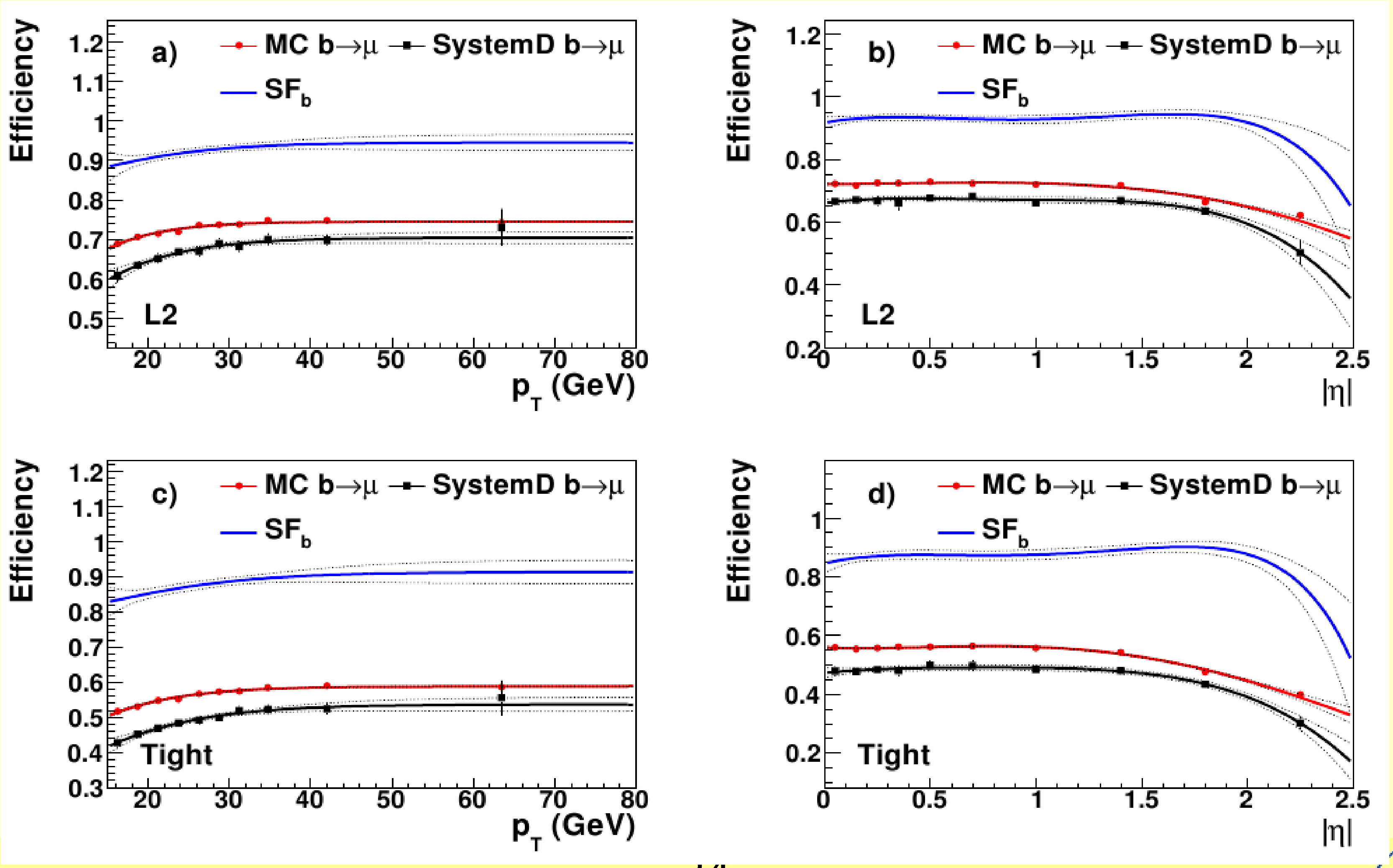
	L2	Tight
Efficiency	65.9%	47.6%
α	0.0%	0.0%
β	0.2%	0.6%
κ_b	0.7%	1.2%
κ_{cl}	0.3%	0.2%
p_T^{rel}	1.0%	0.7%
SystemD Total	1.3%	1.5%

- **B-jet efficiencies errors:** ~2% to ~5%

B-identification efficiency (IX)

Scale factors are measured for 12 operating points

- Optimize efficiency / purity depending on physics channels e.g single / double *(asymmetric)* tags, ...



Sébastien Greder

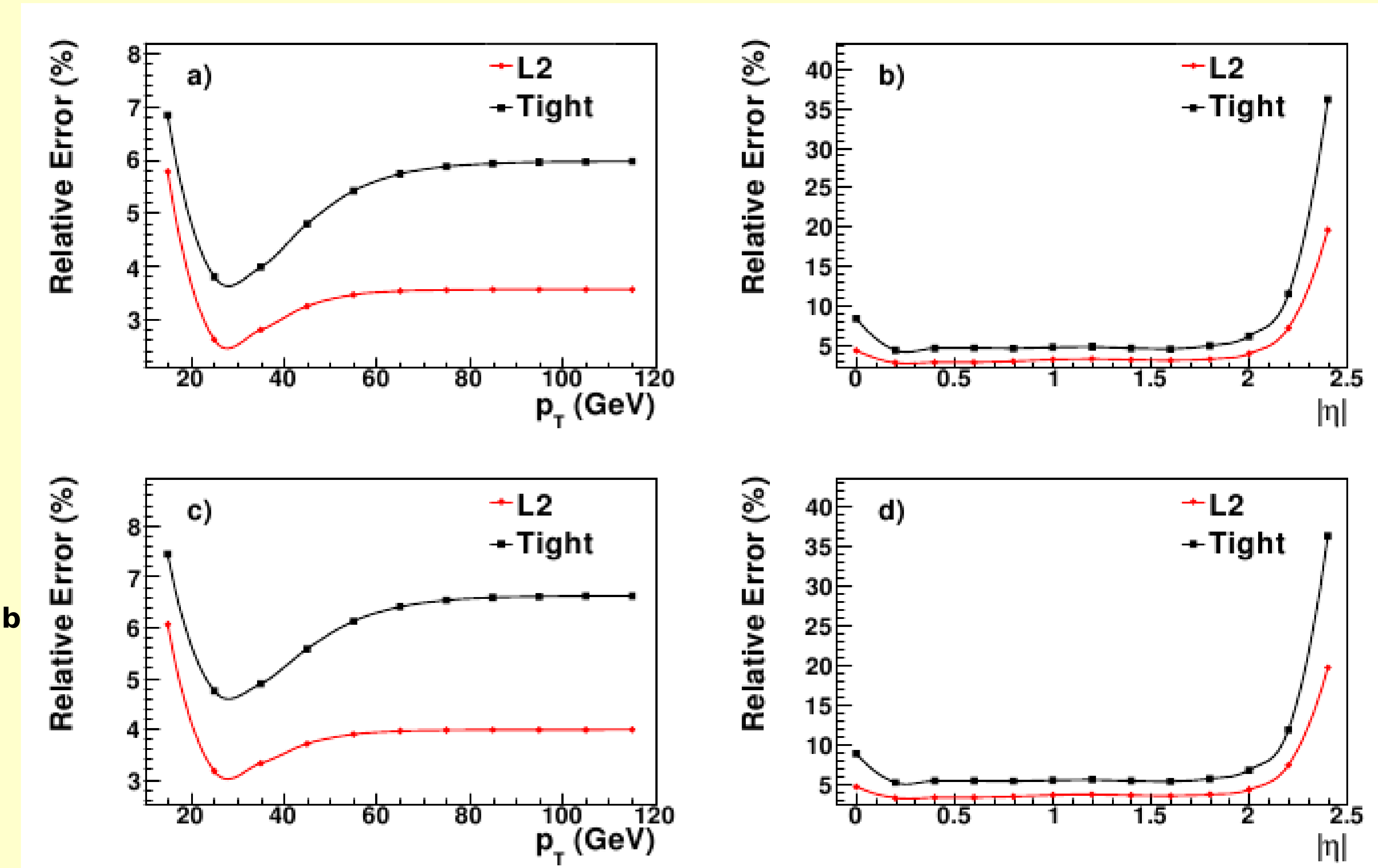
B-identification efficiency (IX)

Scale factors are measured for 12 operating points

- Optimize efficiency / purity depending on physics channels
  e.g single / double *(asymmetric)* tags, ...

SF_b

TRF_b

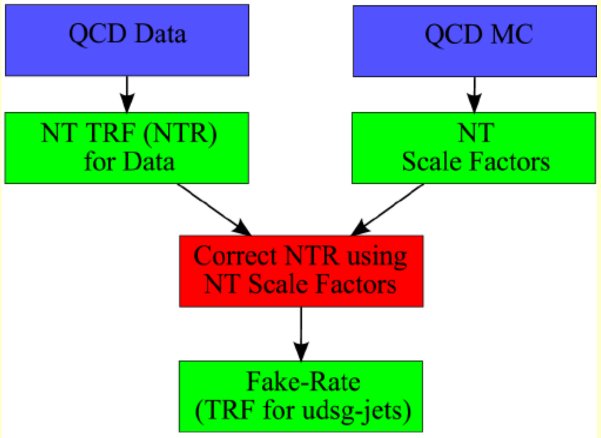


Institut Pluridisciplinaire Hubert Curien STRASBOURG

Fake rate

Goal

- Estimate ε_{light} where light = u,d,s and gluon
- Measured in ***data***



Estimated from *negative tags*

Corrected for:

- HF contamination:
- Neg./Pos. asymmetry:

$$F_{\rm hf} = \varepsilon^{-}_{\rm QCD,light}/\varepsilon^{-}_{\rm QCD,all}$$

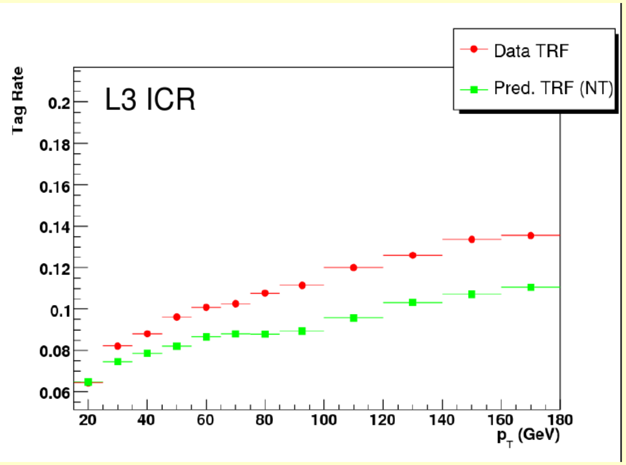
$$F_{\rm lf} = \varepsilon^{+}_{\rm QCD,light}/\varepsilon^{-}_{\rm QCD,light}$$

$$\boxed{\varepsilon_{\rm light} = \varepsilon^{-}_{\rm data} \cdot F_{\rm hf} \cdot F_{\rm lf}}$$

Parameterisation

- F(p_T,η(CC,ICR,EC))

But: NT method **underestimates** fake-rate *(hidden in "experimental" k-factor)*

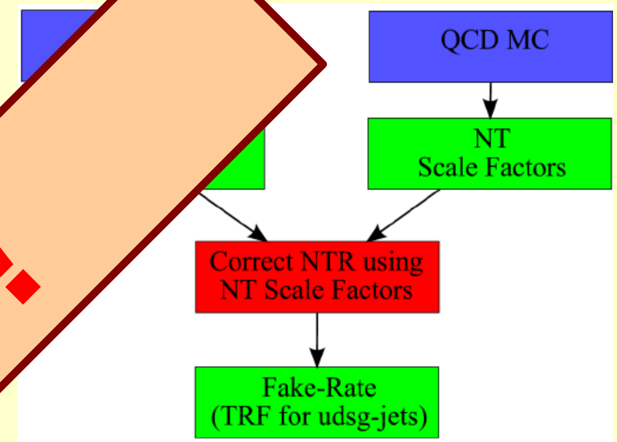


Sébastien Greder

Fake rate

Goal

- Estimate $\varepsilon_{\text{light}}$ where light = u,d,s and gluon
- Measured in ***data***



Estimated from *negative tags*

Corrected for:

- HF contamination:
- Neg./Pos. asymmetry:

$F_{\text{hf}} = \varepsilon^{-}_{\text{Q}}$ … all

$F_{\text{lf}} =$ … CD,light

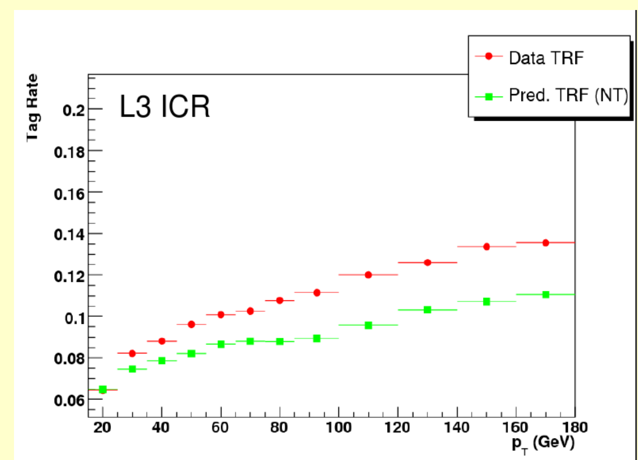
$$\boxed{\varepsilon_{\text{light}} = \varepsilon^{-}_{\text{data}} \cdot F_{\text{hf}} \cdot F_{\text{lf}}}$$

OBSOLETE !

Parameterisation

- F(p_T,η(CC,ICR …

But: NT method un… es fake-rate (*hidden in “experi… ctor)*



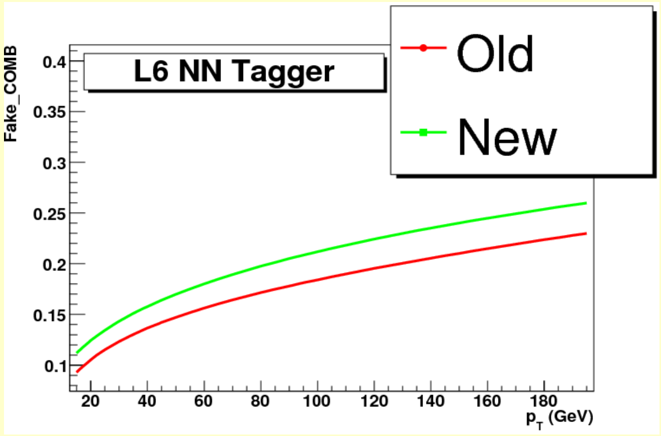
Sébastien Greder

Institut Pluridisciplinaire Hubert Curien STRASBOURG

Fake rate

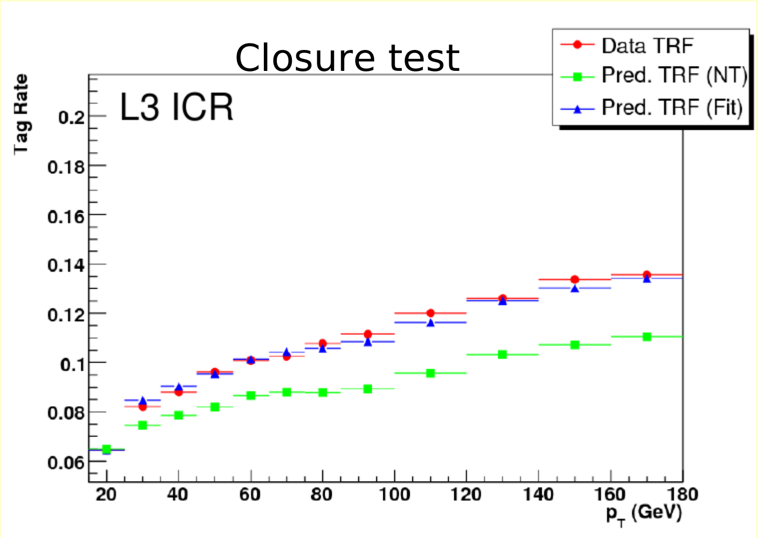
New method (default since summer 2009)

- More data-driven
- Can be applied on specific dataset (e.g hbb)
- Uses b/c tag rate from System8



Results

- Similar shapes
- 20-50% higher rate
- Good closure tests.
- $K^{exp}_{hf} \sim 1$

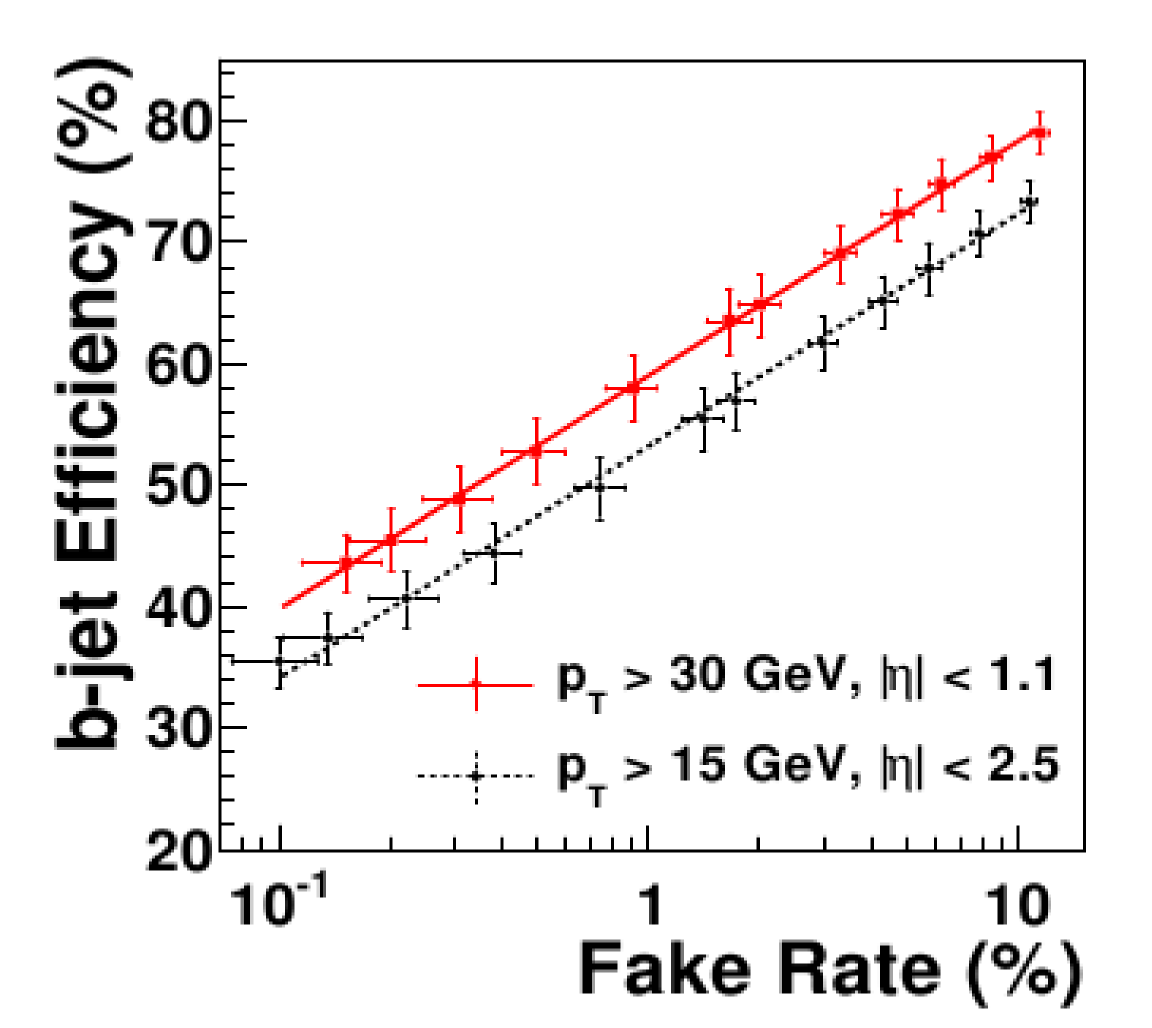


Sébastien Greder

Performance

Final data performance

- Using MC Z decays with data / MC scale factors



Sébastien Greder

Conclusion

- These algorithms have been used in many publications of D0RunII analyses
- SystemD applied to b-identification efficiency measurement is a powerful method with little dependency on simulation
  - *It is already used in both ATLAS and CMS*
- **Other / On-going / Future developments:**
  - MVA taggers
  - Improvements to the algorithms and methods
    - Fake track killer / tracking tuning / neg. tags / ...
    - See talks about Tevatron run extension ... :(
  - Better understanding of detector response
  - But also higher instantaneous luminosity
  - *You can contribute !*

Thank You !

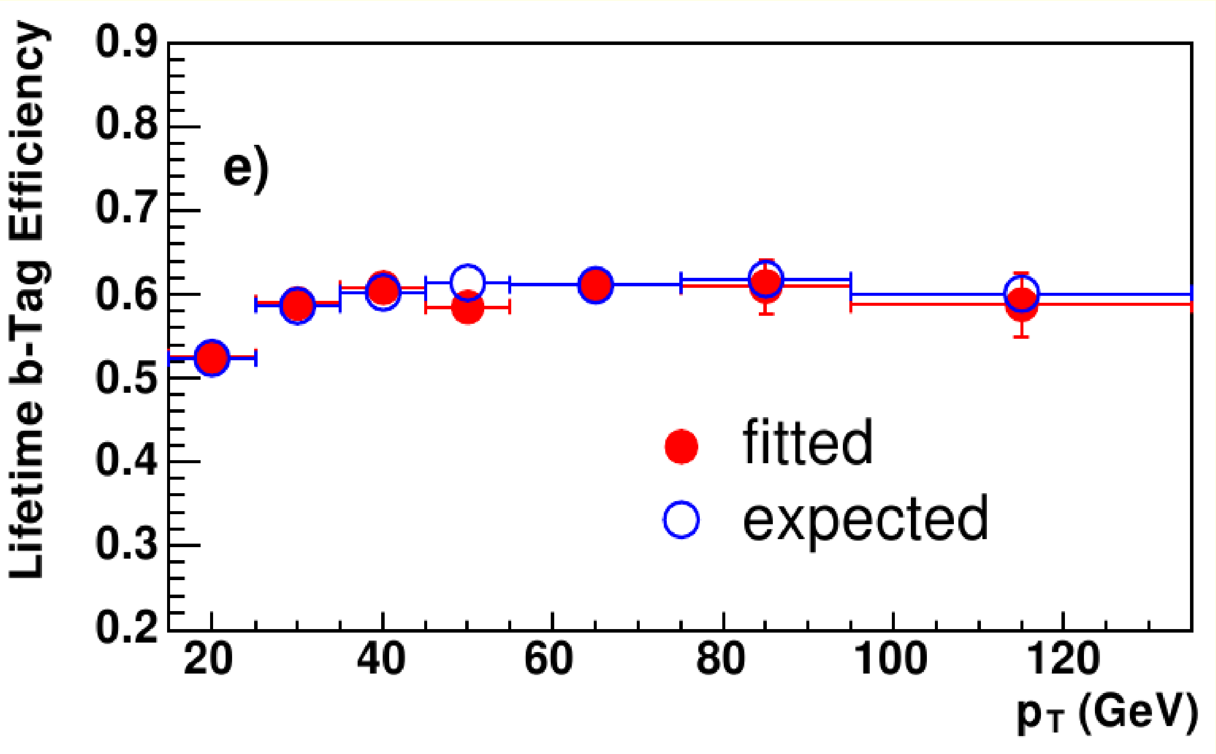
Sébastien Greder

IPHC Institut Pluridisciplinaire Hubert Curien Strasbourg

Back-up

System D in simulated events

Method validation in simulation:



Sébastien Greder